An Information Sources Map for Occupational and Environmental Medicine: Guidance to Network-Based Information Through Domain-Specific Indexing

Scot M. Silverstein, M.D.[1], Perry L. Miller, M.D., Ph.D.[1], Mark R. Cullen, M.D.[2]
[1]Center for Medical Informatics, [2]Occupational and Environmental Medicine Program,
Yale University School of Medicine, New Haven, CT 06510

This paper describes a prototype information sources map (ISM), an on-line information source finder, for Occupational and Environmental Medicine (OEM). The OEM ISM was built as part of the Unified Medical Language System (UMLS) project of the National Library of Medicine. It allows a user to identify sources of on-line information appropriate to a specific OEM question, and connect to the sources. In the OEM ISM we explore a domain-specific method of indexing information source contents, and also a domain-specific user interface. The indexing represents a domain expert's opinion of the specificity of an information source in helping to answer specific types of domain questions. For each information source, an index field represents whether a source might provide useful information in an occupational, industrial, or environmental category. Additional fields represent the degree of specificity of a source in individual question types in each category. The paper discusses the development, design, and implementation of the prototype OEM ISM.

INTRODUCTION

Clinicians and researchers in Occupational and Environmental Medicine (OEM) have a pressing need for easy access to current scientific, technical, and legal information. The information is crucial to effectiveness in diagnosing and treating occupational and environmental health problems, and in advising industrial and environmental personnel. A recent Institute of Medicine report details this need [1].

In recent years there has been rapid growth in the availability of such information. In addition, advances in computer communication are making much of this information available on-line via computer networks, or through telephone modems. On-line access is very helpful, since the information source sites are scattered around the world.

Significant problems still remain, however, which limit timely and widespread use of these on-line information sources. It is difficult, for example, to keep track of, select, and connect to sources that are appropriate to a specific OEM situation or question. This is especially true for those not proficient with computers.

This paper describes a prototype OEM information sources map (ISM), which complements a more general, biomedical-coverage ISM, that we are developing as a step towards solving these problems. Both ISMs are being developed as part of the Unified Medical Language System (UMLS) project of the National Library of Medicine [2].

SAMPLE SESSION

This section describes a sample session with the prototype OEM ISM. In using this program, the user is presented with a series of graphical user interface screens. In the example, the user requests information sources that can help confirm possible occupational causes of a worker's cancer.

The master screen in Figure 1 presents the user with domain-specific choice boxes used to specify the category of the question and its details. To indicate an occupational injury or illness category (referred to in the prototype as "injury") the user selects "**Injured worker evaluation or care.**" The user then selects the "**Causality to injury**" question type to indicate interest in causality information.

The user further constrains the search by entry of a qualifier term. A qualifier is an additional question component, such as a specific injury or hazard. In this example the user specifies a possible work-related neoplasm (cancer). Upon clicking "**Specify injury**", the injury-entry screen in Figure 2 appears. The user selects "**Neoplastic**", and clicks "**Done**" to return to the master screen.

If the user clicks upon "**Specify Source Type**", the source content characteristics can be chosen from a list of seven types. These types are 1) journal citations, 2) book citations, 3) textbooks on-line, 4) other full text on-line, 5) reference database, 6) government document, and 7) directory. The default is all of these. Upon selecting "**Find information**

0195-4210/92/$5.00

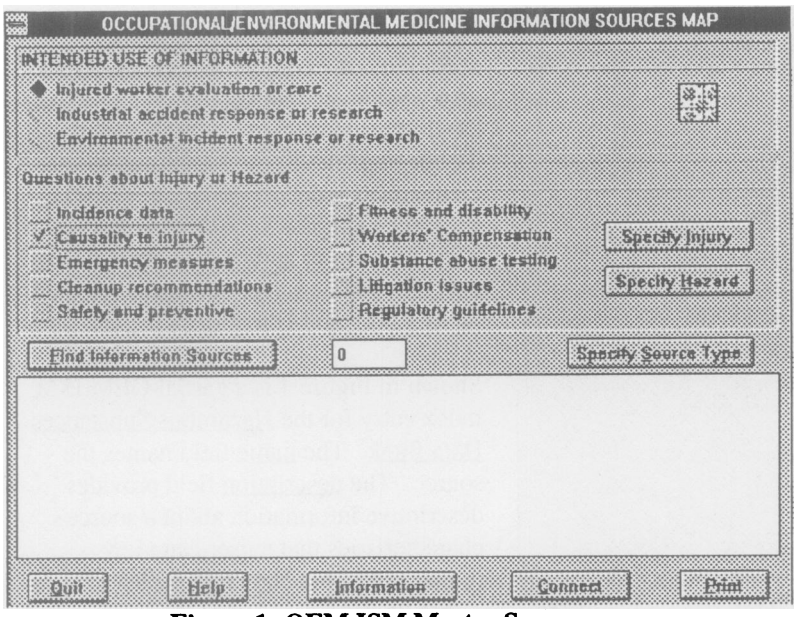

Figure 1: OEM ISM Master Screen

sources", the user is presented with a list of information sources, grouped in three categories of specificity. The list is presented in the large white results box at the bottom of the master screen. If the list is too large for this space, the user may scroll through it by clicking on a "scroll bar" icon. For clarity, only the contents of the results box, with information source names grouped in three categories of specificity, is shown in Figure 3. (An explanation of these categories is presented later.)

The user may now click on the name of a particular source in the list and bring up a detailed description of its characteristics by clicking on "**Information.**" This is helpful in further deciding if an information source meets the user's needs. The user may also initiate a connection to the information source without knowledge of connection protocols by clicking on "**Connect**". A direct connection will be established with the information source. The user's computer acts as a communications terminal until disconnection, at which time the OEM ISM returns.

As illustrated, this ISM is designed to allow even a novice user to be guided to, and to review, the potential usefulness of sources of OEM information based on specific needs. Access to this information might otherwise have been difficult. This working prototype currently includes 31 information sources. It can potentially be expanded to include a much greater number of sources. A research prototype contains 103 information sources, as described below.

At this stage of development in the prototype ISM, the user must query an information source manually after connection. This requires knowledge of an information source's search techniques. There is, as yet, no standardization of search techniques, which vary from source to source. Ultimately, this approach might be enhanced to include the ability to search the sources themselves in a facilitated fashion.

PROGRAM DESIGN

The OEM ISM is implemented using a client-server model. This allows the database of information sources to be stored and maintained centrally. Two computer programs are required to implement this model. A client program runs on an IBM personal computer or compatible running Microsoft Windows and interacts with the user. This ISM client communicates with a server computer using the DynaComm communications package [3]. The server is a Sun SparcStation-2 UNIX workstation. A server program, currently written in Lisp, processes user requests for on-line OEM information sources.

Selection of Useful OEM Information Sources

The sources included in the ISM must contain specific technical information which is needed by practitioners and researchers of OEM, as recommended by the Institute of Medicine. Timely access is needed, at minimum, to information about 1) hazardous substances, 2) the risk of exposure causing injury or illness, 3) expert advice about diagnosis and medical management of persons exposed to hazardous substances, 4) current information about hazardous substances produced by local industry, 5) geographic patterns of relevant clinical illness, 6) governmental case reporting requirements, and 7) related legal matters.

Potential on-line information sources covering these topics were selected from the UMLS Information Sources Map set of information sources [4] and the Directory of On-Line Databases [5].

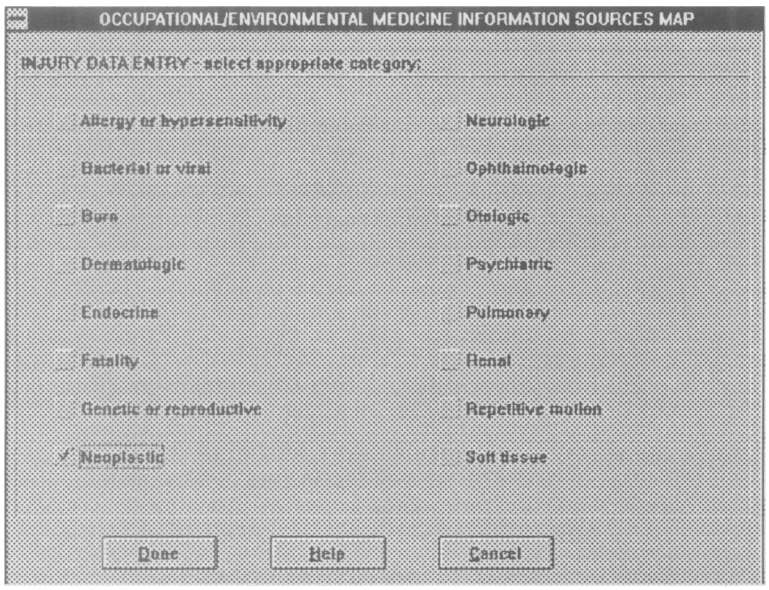

Figure 2: Injury Entry Screen

The information sources are diverse in content and also in underlying composition, ranging from journal citations and textbooks, to collections of raw data gathered by hospitals and governmental agencies.

The OEM ISM working prototype contains 31 sources, a subset of the UMLS Information Sources Map set of sources. This version is a fully operational prototype. A second version, an OEM ISM research prototype, contains these 31 sources plus an additional 72 sources not in the UMLS ISM set. The 72 additional information sources currently lack full descriptive information and source connection ability. The research prototype with its total of 103 information sources has been used to more fully test the usefulness of the OEM indexing in making information source recommendations.

=== Focused information source relevant to query:
CANCERLIT
Chemical Carcinogenesis Research Information System
GENE-TOX
Registry of Toxic Effects of Chemical Substances
Sigma-Aldrich Material Safety Data Sheets
TOXLINE

=== General information source that may be relevant:
Comprehensive Core Medical Library
Current Awareness in Biological Sciences
Current Contents Search
MEDLINE
Occupational Safety and Health (NIOSH)

=== Reminder of source that merits consideration:
CA Search

Figure 3: Results Box

Information Source Indexing

Methods were needed to allow retrieval of appropriate sources. We particularly wanted to allow OEM experts to encode their opinions regarding the content and relative specificity of information sources. We also wanted the indexing system to yield explicit responses to the OEM questions. To accomplish this, new field types were created and added to the existing UMLS ISM information sources indexing system. The inclusion of additional data elements useful in source selection was anticipated by Lindberg and Humphreys [6].

Shown in Figure 4 is a partial OEM ISM index entry for the Hazardous Substances Data Bank. The name field names the source. The description field provides descriptive information about a source's characteristics that a user can view. (Other fields, not shown, provide information about human contacts at the source site, cost, how to connect to the source, and other technical matters.)

The OEM category field encodes a domain expert's opinion of the coverage of the source in the three broad "Intended use of information" categories of the OEM ISM master screen (injured worker, industrial accident, and environmental incident).

The OEM content field encodes the specific questions and qualifiers for which a domain expert feels the source might contain answers. The first two items in this field represent a question and qualifier pair which the content of this information source may address, such as 1) safety issues regarding 2) chemical hazards. This index field can also contain a "wild card" to reduce the length of an information source index. For example, if a source's content might be relevant in safety issues for any hazard, the index entry would contain "**safety hazard_any.**"

The third item in the OEM content field represents the domain expert's opinion as to the specificity of the information source in providing answers to this particular question. In this prototype, there are three categories of specificity of a source in answering a question, as follows:

1) "Focused information source relevant to query" means a specialized source, covering a topic closely associated with the question. For example, the AIDS Knowledge Base is a focused source useful for questions about the incidence of AIDS.

2) "General information source that may be relevant" means a broad-coverage source that is likely to have useful information. For example, the Comprehensive Core Medical Library is a general source useful for questions about emergency treatment of chemical exposure.
3) "Reminder of a source that merits consideration" means a source that is less likely to have focused information closely related to the question, but should be evaluated in a thorough review.

Name	Hazardous Substances Data Bank		
Description	(Descriptive text about the information source)		
OEM category	Injury | Industrial | Environmental		
OEM content	safety	hazard_chemical	focused |
	causality	injury_neoplasm	focused |
	causality	injury_allergy	focused |
	causality	injury_neurologic	reminder |
	regulatory	hazard_chemical	focused

Figure 4: Indexing in OEM ISM

Indexing must be done by individuals with considerable domain knowledge. This type of indexing is only practical in a limited domain with focused information needs such as OEM. The indexing process is not complex and should be readily understood by non-computer-proficient domain experts. Indexing is not overly time-consuming when performed in the focused domain of OEM. We believe that the approach should be extensible to more complex queries.

User Interface
The scope of the domain was an important factor in the development of an interface which is "friendly" to OEM users. OEM is a relatively well-defined domain compared to broader fields such as Internal Medicine, and has more focused information needs. It was therefore felt that a menu-based approach to question entry could cover many of the needs without use of unrestricted, user-typed search words. The advantage of a menu-based approach is that questions can be framed in the precise language of the domain. This should make the interface easier to use.

This approach would not be practical in a very broad domain, although we can envision the creation of analogous menu-based applications for other relatively constrained biomedical domains. A disadvantage of the approach is the limitation of the number of question types and qualifiers. It is encouraging that the current small prototype seems able to cover a useful portion of OEM needs. An increase in the number of questions and/or qualifier types can be done by increasing the number of screens, or by adding "scrolling lists" of terms from which the user can make selections.

Client-Server Interaction and Server Operation
The database of information sources resides on a UNIX-based server computer which is itself accessed on-line. The list of sources and its contents index is maintained centrally. The client computer program is written in the commercial communications scripting language DynaComm. The client program typically resides on a networked file server, for loading into the user's PC when a search is to be done.

The client program then runs on the user's machine and collects the components of a user's questions. It sends these to the server as text strings. For example, in the "Intended use of information" category Injured worker evaluation or care a question about safety issues regarding chemical hazards causes the strings "**injury**", "**safety**" and "**hazard_chemical**" to be sent to the server.

The server program is written in Lisp in our prototype. In evaluating whether to return a source to the client as potentially useful, the server checks for a match between the user question strings received from the client, and entries in the information source index (OEM category and OEM content fields). If a match is found, it will return the name of the source and the specificity ("focused", "general", or "reminder") stored as the third item in each OEM content field. The client displays the list of returned source names, in groups based on the categories of specificity.

When a user clicks on a source name and then clicks "**Connect**", the user's computer will connect directly to the source. No further intervention is needed by the user for this to occur. DynaComm handles the details by executing connection scripts. These connection scripts are also maintained and updated centrally on a file server. Once connected to a source, a user may conduct a search. As mentioned previously, the user must be familiar with the searching methods of each source. Facilitation of the actual search by automated means could be a future development.

Partitioning of Information Sources by Indexing
Indexing of the information sources should partition the sources into relatively limited subsets in order to be useful. We performed a preliminary evaluation of

partitioning of the information sources into small subsets by the 10 general question types. The number of sources retrieved for a question type changes as different "Intended use of information" categories (injury, industrial, or environmental) and qualifier terms (injury or hazard types) are chosen. A summary of this partitioning is shown in Table 1.

The results are shown in the form **"average (least - most)."** Average indicates the rounded average of 27 values, which are the numbers of information sources retrieved for a question type with each of 27 possible qualifier terms (injury or hazard). The average is followed in parentheses by the least (minimum) and most (maximum) number of sources retrieved for a question type with individual qualifiers.

For example, in the "Intended use of information" category **Injury**, the question type "emergency measures" entered with each of 27 possible injuries and hazards retrieves a varying number of sources, averaging 8 sources. When the hazard entered with this question is ionizing radiation, the least number of sources for any of the qualifiers is retrieved (7 sources). When the hazard entered is chemicals, the most number of sources is retrieved (14 sources). Hence, in Table 1 the numbers **"8 (7-14)"** appear in row 3 (emergency measures), in the column labeled "injury."

Table 1: Summary of Partitioning

Question type	Intended Use of Information		
	Injury	Industrial	Environmental
1. Incidence	16 (13-20)	13 (12-16)	6 (5-8)
2. Causality	17 (9-33)	15 (8-26)	14 (5-25)
3. Emergency measures	8 (7-14)	8 (7-14)	5 (4-8)
4. Cleanup	4 (2-11)	5 (2-15)	5 (1-17)
5. Safety	12 (10-27)	11 (8-27)	7 (2-24)
6. Fitness	4 (2-7)	3 (2-5)	1 (1-2)
7. Worker's compensation	5 (5-7)	5 (5-5)	1 (1-1)
8. Substance abuse	11 (11-11)	10 (10-10)	7 (7-7)
9. Litigation	11 (10-15)	10 (9-14)	5 (4-9)
10. Regulatory	11 (9-20)	10 (8-19)	6 (3-18)

These results suggest that the indexing provides a useful partitioning of the information sources, in the sense that a specific query will retrieve a focused subset of the information sources. The information source type selectors (e.g., journal article, textbook on-line, etc.) partition the sources further and cause the retrieval of even smaller subsets.

CONCLUSIONS

The "domain-specific" indexing approach used in the prototype OEM ISM indexes sources not by subject heading or key words, but by an interpretation of content relevant to specific domain questions and qualifiers. The interpretation is performed by domain experts. A domain expert's knowledge is encoded explicitly in an attempt to give focus and direction to information source recommendations. We believe that this approach partitions OEM information sources with overlapping coverage in a potentially useful manner relative to the OEM professional's information needs.

The user interface of the OEM ISM is also domain-specific. We believe this makes the interface easier to use, and helps minimize confusion and wasted interaction by the user. The prototype represents an initial step in the construction of a focused ISM that helps meet the needs of a specific clinical domain. It explores a design which is focused on the actual question types that OEM practitioners are likely to ask.

ACKNOWLEDGEMENTS

This research was supported in part by NIH Grant T15 LM07056 and contract N01 LM13537 from the National Library of Medicine, and by an equipment grant from Sun Microsystems, Inc.

Reference

[1]. Institute of Medicine, Meeting Physician's Needs for Medical Information on Occupations and Environments. Report of a Study. National Academy Press, 1990.

[2]. P.L. Miller, L.W. Wright, S.J. Frawley, J.I. Clyman, S.M. Powsner. Selecting relevant information resources in a network-based environment: The UMLS information sources map. MEDINFO-92, Elsevier Science Publishers B.V., 1512-1517, Sept. 1992.

[3]. Dynacomm 3.1, Future Soft Engineering, Inc., Houston, Texas

[4]. D.R. Masys, B.L. Humphreys. Structure and Function of the UMLS Information Sources Map. MEDINFO-92, 1518-1521, Sept. 1992.

[5]. Directory of On-Line Databases. Cuadra/Elsevier, NY, NY, Jan. 1992.

[6]. B.L. Humphreys, D.A.B. Lindberg. The Unified Medical Language System Project: A Distributed Experiment in Improving Access to Biomedical Information. MEDINFO-92, 1499, Sept. 1992.